Going GREAN:
A Novel Framework and Evaluation Metric for the Graph-to-Text Generation Task

Oren Sheffer [1] **Or Castel** [1] **Raz Landau** [1]

[1] Tel Aviv University
{orensheffer,orcastel,razlandau}@mail.tau.ac.il

Abstract

Current models for generating text from a Knowledge Graph (KG) rely heavily on a reference-text for both training and evaluating. While this enables training models to generate valid texts, we claim that it restricts the models in term of results quality and relevant data sets scale and quality. We attempt to take the first steps towards eliminating this restriction, and propose two new pillar for a novel reference-text-free framework for the Graph-to-Text task: 1. GREAN, a new automatic evaluation metric based on extracting a KG from the generated text 2. Performing transfer learning from a pre-trained language generative model when training a model for the task. We show that GREAN outperforms current reference-text-based automatic metrics for assessing factual accuracy through a human evaluation study and that leveraging a pre-trained generative model yields promising results for generating valid texts without a reference text used during model training. These serve as a proof of concept for the reference-text-free future framework. [1]

1 Intro

One of the classic problems in natural language generation (NLG) involves taking structured data, such as a graph, as input, and producing text that adequately and fluently describes this data as output. This task has seen a growth in interest over the past few years, with most recent models using attention-based encoder-decoder architecture (Gardent et al. 2017b, Marcheggiani & Perez-Beltrachini 2018, Koncel-Kedziorski et al. 2019).

All these models share a characteristic in common - their loss function during training is computed in relation to the next-word prediction in the reference text. Moreover, the only gold standard metric For text generation tasks is to show the output to humans for judging its quality, but this is too expensive to apply repeatedly anytime small modifications are made to a system. Hence, automatic metrics that compare the generated text to one or more reference texts are routinely used.

The fact that the both the learning and evaluation process is heavily dependent on the structure of the data - the existence of a reference text - forces using only datasets that has both KG and text for each sample in the data. This constraint comes with a heavy cost. Manually curated datasets, where a text is manually curated to match a KG, or a KG is manually curated to match a given text (Gardent et al. 2017a, Novikova et al. 2017, Sun et al. 2018) requires laboursome, sensitive and sometimes even domain-specific annotation process. On the other hand, when datasets are collected automatically and heuristically (Wiseman et al. 2017, Lebret et al. 2016), inaccuracies between the KG and the facts that are truly represented in the reference text are frequent and lead to erroneous results (Dhingra et al., 2019).

We attempt to eliminate the dependence on reference texts, focusing first on the evaluation process, by developing a new metric, GREAN (**GR**aph **E**xtr**A**ctio**N**), that uses an Information Extraction (IE) model to create a KG representing the generated text and compare it to the reference KG. We show that GREAN improves correlation with human judgments over popular automatic metrics for assessing the factual accuracy of generated text. Figure 1 shows an example GREAN's strength.

The missing piece of the puzzle is the designing of a loss function and model training process that is not dependent on predicting the reference text, but rather on comparing facts represented in the generated text and those in the KG. Without the loss function of next-word-prediction, the model needs the ability to be trained to generate valid language.

[1] Data and code are available at https://github.com/ocastel/txt2graph

Reference:	**In this paper, we introduce GREAN, a reference-text-free evaluation metric for the Graph-to-Text task**	***BLEU***	***ROUGE***	***GREAN***
Generated 1:	In this paper, we introduce PURPLE, a reference-text-free evaluation metric for the Graph-to-Text task	0.782	0.944	0
Generated 2:	Compared to other automatic metrics, GREAN is a reference-text-free evaluation metric for the Graph-to-Text task based on graph extraction	0.366	0.598	1

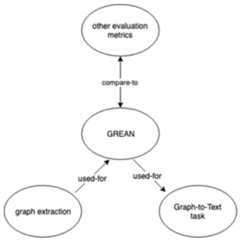


Figure 1. A reference text, its KG (right) and 2 hypothetical generated texts with scores assigned to them by automatic evaluation metrics

To tackle this, we present a first attempt, to our knowledge, to perform transfer learning from a pre-trained language generative model for the Graph-to-Text task . Using such model should enable the generation of valid texts even without the next-word-prediction training process. Since using the reference text as a target for the model is what enables it to generate valid texts, we believe that such approach is necessary.

Another advantage of using transfer learning rather than only reference text as a target is the limitation of corpus - a model wouldn't support words that didn't appear in the reference texts, with the exception of models that uses copy mechanisms - and even those would be limited to words that appear in the KG.

Given that a reference text will no longer be a necessity at any step of the process, we propose the GREAN Framework (named after the evaluation metric at its core) for the Graph-to-Text task, that relies on a pre-trained IE model:

1. Datasets are created with minimal cost and resources using such model from any corpus, as long as the performance of the IE model on the corpus type is good. They can also be randomly generated as long as the relation types are supported by the IE model.
2. Natural language structures and texts are inducted using transfer learning from a pre-trained generative model, rather from relying on predicting a reference text.
3. Evaluation, one that truly measures the task performance of a model, is done using GREAN evaluation metric that is based on the IE model used to generate the dataset.

To this end, we make the following contributions:

- We introduce a reference-text-free metric, GREAN, to analyze the factual accuracy of generated text and compare it against traditional reference-text-based metrics.
- We perform transfer learning in the context of the Graph-to-Text task, in a way that is easily applicable to commonly used model design.
- We conduct experiments to compare our proposed metric against human evaluation of factual accuracy of generated text and show that it is better correlated with human judgment when compared to traditional metrics.

2 Related Work

Despite the popularity of the Graph-to-Text Generation task, almost all current work heavily relies on reference texts for both training and evaluating.

Most of recent models designed for Graph-to-Text or similar Structured Data-to-Text are based on the attention-based encoder-decoder architecture, and differ in application of a copy mechanism and the handling of the graph input.

Although most previous known models, Koncel-Kedziorski et al. (2019), Gardent et al. (2017b), Marcheggiani & Perez-Beltrachini (2018), Wiseman et al. (2017), Veličković et al. (2017) - to name a few- differ in the pre-processing done to the KG, its representation in the encoder process, type of encoder applied and mechanisms added to the decoder, they all use a loss function of next-word prediction in the reference text.

In addition, we note that although text generation models have become popular lately for their capability to generate natural language of many domains - such are GPT-2 (Radford et al., 2019) and XLNet (Yang et al., 2019), and the rising popularity and effectiveness of transfer learning in NLP tasks - none of the NLG-specific models mentioned above take any advantage of text generative models.

As for the evaluation process, it is common to classify the existing metrics either as extrinsic metrics - those that are specific to tasks, or intrinsic metrics like grammatically, that are based on the analysis of the text.

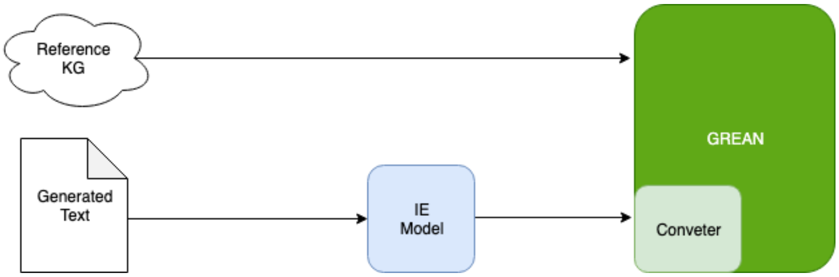


Figure 2. Illustration of GREAN's architecture

Usually, in text-generation a popular machine translation metric is used in comparison to a reference text - such as BLEU, ROUGE or METEOR. These metrics, however, rewards highly text that is natural and fluent rather than factual. A few articles along the past few years tried to tackle this issue in various ways.

Wiseman et al. (2017) were the first to suggest using information extraction to factually compare between the reference KG and the generated text. However, their work was based on tabular information and was restricted to the very specific domain of basketball games. We propose a method of evaluation that could be generic to any kind of classic knowledge base, and in addition will be domain generic.

Dhingra et al. (2019) noted the disadvantage of current evaluation metrics in cases where the reference text ”hallucinates” information and diverge from the reference data. As mentioned, this is an known problem when dealing with any automatically and heuristically collected datasets. However, their proposed metric, PARENT, still rely on the reference text for evaluation (n-gram alignment), such as the previous popular metrics available.

Goodrich et al. (2019) $fact_{acc}$ metric for text summarizing task most resembles ours, estimating the factual accuracy (precision) of generated text by extracting facts from it. The main difference is them focusing only on a mteric similar to precision, neglecting recall performance and ”hallucinations”, both in the automatic evaluation and in the way human evaluation was conducted. While they compare a generated text with a reference text, in our human evaluation the comparison is between the generated text and the reference KG.

3 GREAN

To evaluate a Graph-to-Text generation model, GREAN evaluates each instance $(\mathsf{Ref}_{\mathsf{KG}}, \mathsf{Gen}_{\mathsf{text}})$ separately, by extracting $\mathsf{Gen}_{\mathsf{KG}}$ from $\mathsf{Gen}_{\mathsf{text}}$ and computing the precision and recall of $\mathsf{Gen}_{\mathsf{KG}}$ against $\mathsf{Ref}_{\mathsf{KG}}$. Formally, let:

$$\mathrm{GREAN}_p = \frac{|\mathsf{Ref}_{\mathsf{KG}} \cap \mathsf{Gen}_{\mathsf{KG}}|}{|\mathsf{Gen}_{\mathsf{KG}}|}$$

$$\mathrm{GREAN}_r = \frac{|\mathsf{Ref}_{\mathsf{KG}} \cap \mathsf{Gen}_{\mathsf{KG}}|}{|\mathsf{Ref}_{\mathsf{KG}}|}$$

We can then define GREAN as the F1 score between $\mathsf{Gen}_{\mathsf{KG}}$ and $\mathsf{Ref}_{\mathsf{KG}}$, i.e:

$$\mathrm{GREAN} = 2 \cdot \frac{\mathrm{GREAN}_p \cdot \mathrm{GREAN}_r}{\mathrm{GREAN}_p + \mathrm{GREAN}_r}$$

For example, consider reference KG : {*(GREAN, compare-to, BLEU), (GREAN, evaluate-for, Graph-to-Text Generation task)*} and generated text: *GREAN is a new evaluation metric for the Graph-to-Text Generation task.* Then, for $\mathsf{Gen}_{\mathsf{KG}}$ ={*(GREAN, evaluate-for, Graph-to-Text Generation task)*} the metric GREAN = $\frac{1 \cdot 0.5}{1+0.5} = 0.33$ indicates there is partial factual consistency in the generated text.

An illustration of the architecture is shown in Figure 2.

4 Transfer Learning

We propose a simple and generic method for incorporating a pre-trained RNN language generation model into most KG-based generative models. The method requires the KG-based model to use at least the same vocabulary as the pre-trained one - otherwise, we would have to carefully treat missing words which might harm the new model computationally, and limit its potential output.

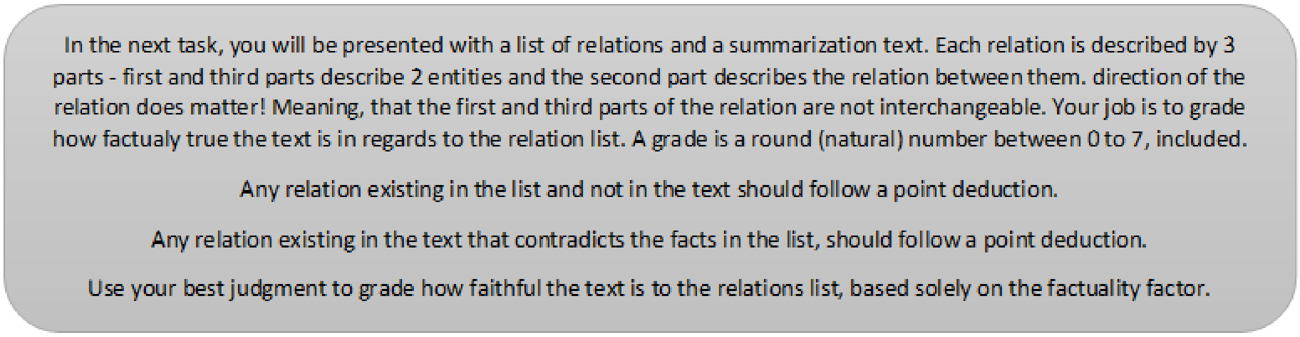
$$\alpha = \sigma(W^{copy} \times h_t^{gen} + b^{copy}) \quad (1)$$

$$out_t^{gen} = softmax(W^{gen} \times h_t^{gen}) \quad (2)$$

$$out_t^{pre} = softmax(W^{pre} \times h_t^{pre}) \quad (3)$$

$$out_t = \alpha \cdot out_t^{gen} + (1 - \alpha) \cdot out_t^{pre} \quad (4)$$

The integration is done on the decoder step, where most models decode using input-feeding decoder (Luong et al.,



In the next task, you will be presented with a list of relations and a summarization text. Each relation is described by 3 parts - first and third parts describe 2 entities and the second part describes the relation between them. direction of the relation does matter! Meaning, that the first and third parts of the relation are not interchangeable. Your job is to grade how factualy true the text is in regards to the relation list. A grade is a round (natural) number between 0 to 7, included.

Any relation existing in the list and not in the text should follow a point deduction.

Any relation existing in the text that contradicts the facts in the list, should follow a point deduction.

Use your best judgment to grade how faithful the text is to the relations list, based solely on the factuality factor.

Figure 3. A screenshot of the interface presented to human evaluators to judge the factual accuracies of generated text

2015). At time t the decoder emits h_t^{gen} and c_t^{gen}. The output h_t^{gen} is used in order to determine a ”copy” coefficient α (equation 1). The pre-trained model is fed with the thus-far selected sequence of words and emits it's own h_t^{pre} and c_t^{pre}. c_t^{pre} is used as context for the pre-trained model on it's next step. Then, the two vectors h_t^{gen} and h_t^{pre} are scaled using W^{gen} and W^{pre} respectively, to the size of the vocabulary. Softmax is applied on both results which are then summed, using α as a weighting factor (equation 4). This output is then used for next-token distribution by the model decoder.

During learning, there are two approaches that can be taken: The first, is to prevent the learner from altering the weights of the pre-trained generative model, thus using it as a copy source that affects distribution over the whole corpus. The second is to allow such alteration, which is the approach generally taken when performing transfer learning.

5 Experiments & Results

We have experimented both with the proposed GREAN evaluation metric and with extending existing model with transfer learning from a pre-trained language model.
First, we describe the dataset and model used throughout our experiments. Following, we describe our first experiment, where we show the effectiveness of our proposed metric on judging the factual accuracy of generated text. Then we describe the methodology used to compare human judgment of factual accuracy and how we compare our metric against that baseline. Next, we describe our experiments with applying transfer learning with pre-trained language generation model into an already existing model for text generation from a KG.

5.1 Data and Models

Our main experiments are on the the AGENDA dataset (Koncel-Kedziorski et al., 2019), where each sample is made of scientific abstracts and a KG extracted from the reference text with Scientific Information Extractor (SciIE) Luan et al. (2018) SciIE is a multi-task learning setup of which all three tasks of entity recognition, relation extraction, and coreference resolution are treated as multinomial classification problems with shared span representations. We also used SciIE for extracting relations from the generated texts, using the same weights that were used for constructing AGENDA data set originally.

The model we use both for comparing the GREAN evaluations of texts generated from the AGENDA data set and experiment with transfer learning is GraphWriter, Koncel-Kedziorski et al. (2019), which is based on Veličković et al. (2017) GraphAttention Network (GAT). The GAT model introduced a multi-head attention mechanism over the graphs neighborhood in the encoding phase, and can be regarded as a masked version of an attention mechanism. GraphWriter uses the concepts introduced in GAT architecture, and adds on top of every graph-attention unit some non-linearity and FC layers. Its decoder also uses, in addition to attention over the encoding, a copy mechanism directly from the graph. A pre-processing step is done prior to training, where the directed graph is transformed into an indirect graph without loss of information, so it would better fit the graph-attention units.

5.2 GREAN Evaluation Metric

5.2.1 Human Evaluation

Because our dataset is scientific in nature, evaluations must be done by experts and we can only collect a limited number of these high quality datapoints. The study was conducted by 3 experts (i.e. computer science graduate students) who were familiar with the abstract writing task and the content of the abstracts they judged.

Each evaluator was shown 100 tables with a generated text and the reference KG, was asked to rate the generated text and score it from 0-7 with 7 being highest factual accuracy. Figure 3 shows the interface a human evaluator uses in our experiment. Finally, the median ranking was taken for each example, resulting in every text having a natural score between 0 and 7.

5.2.2 Compared Metrics

We compare commonly used text-based automatic metrics including BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) - both are variants of n-gram overlap measure that are popular in text generation and machine translation tasks, and METEOR (Denkowski and Lavie, 2014) - a machine translation with paraphrase and language-specific considerations based on the harmonic mean of unigram precision and recall - all using their publicly available implementations.

METRIC	PEARSON	SPEARMAN
BLEU	0.074	-0.005
ROUGE	0.2	0.165
METEOR	0.167	0.123
GREAN	0.687	0.696

Table 1. Correlation of metrics with human judgments

5.2.3 Results

A comparison of all automatic metrics in terms of correlation to human judgment is shown in Table 1. Our proposed metric GREAN outperforms all other evaluation metrics. We can see that there is in fact *no* correlation between BLEU and the human judgment regarding the factuality of the text, which only emphasizes GREAN's importance. Even automatic metrics that are known to be correlated with human judgment such as ROUGE, are shown to have much less correlation than GREAN. Similar trends can be observed for both Spearman and Pearson correlations, strengthening our confidence in these results.

5.3 Transfer Learning

5.3.1 Pre-Trained Generative Model

For a pre-trained language model we trained a single layered LSTM (Hochreiter & Schmidhuber, 1997) with hidden states and embedding dimensions fixed at 500. The corpus on which the model was trained is the reference texts in the AGENDA data set - which promises the model would rely on the same vocabulary GraphWriter later relies on. The model was trained with negative joint-log likelihood of the target text.

5.3.2 Modeling

We added the pre-trained generative model to the architecture of the GraphWriter model, as described above. GraphWriter applies a copy mechanism from the graph nodes after calculating the vocabulary distribution for the next word. Thus we added the transfer mechanism following the vocabulary distribution calculation and prior to calculating and applying the copy mechanism from the graph itself. We use GraphWriter in its default form - hidden states and embedding dimensions are fixed to 500 and attentions learn 500 dimensional projections. All other parameters are as described in (Koncel-Kedziorski et al., 2019). We ran two versions of training, as described in ”Data and Models” section - one where the pre-trained model weights are not learnable - ”GraphWriterTF-Freeze”, thus the mechanism is purely a ”copy” mechanism, and the other where the weights are learnable - ”GraphWriterTF-NoFreeze”. Each model was trained for 15 epochs with a fixed learning rate of 0.1.

5.3.3 Results

A comparison between the original GraphWriter architecture result, as used in the first experiment of BLEU, ROUGE and GREAN metrics are shown in Table 2.

MODEL	BLEU	ROUGE
GRAPHWRITER	0.117	0.274
GRAPHWRITERTF-FREEZE	0.093	0.234
GRAPHWRITERTF-NOFREEZE	0.116	0.267

Table 2. Average model scores on AGENDA dataset

While not producing any state of the art results, the ability to preform as well as other Graph-to-Text models such as GraphWriter can be seen as a definite proof of concept. The model, built upon a pre-trained generative model - achieves similar BLEU and ROUGE scores. These results are a step towards the ability to build an equivalent model without relying at all on reference texts. It is worth mentioning that tracking the α coefficient suggests that the models were using the pre-trained sub-model at all stages of training.

6 Conclusions and Future Work

This work tackled two different but somewhat related issues seen in most recent state of the art models for the Graph-to-Text task. More specifically, we tried to make a step towards creation of a framework for generating text that is factually based on a knowledge base, with human-like text fluency and variety, without using a reference text in order to create it. Recent SOTA models both learn and are evaluated on metrics that are related to a reference text that is attached to the reference knowledge base, and are focused on creating a human-like text.

We presented GREAN, a new first of its kind framework for training and evaluating a Graph-to-Text models based on information extraction. We showed that GREAN's evaluation metric achieves noticeable better results in terms of human evaluation of factuality, compared to the automatic metric used commonly for this task. We also presented a technique for performing transfer learning from pre-trained generative model, that can later be used when training models with a loss function independent from a reference text.

We conclude that this work is a first significant step towards creating a reference-text-free Graph-to-Text framework and models.

Future work on this task would be the design of a loss function and learning process that will not be dependent on the reference text while better representing the factuality presented in the text compared to the KG. Such a factuality based loss goes beyond the next-word loss and should be able to capture information spread over long dependencies in the sentence. A factuality based criteria might not be differentiable, thus cannot be optimized using the usual backpropogation. To deal with the non differeatiable criteria, we plan to design a factuality based reward function and borrow Reinforcement Learning based method (policy gradient) that allow for optimization of such non differentiable reward functions.

References

Dhingra, B., Faruqui, M., Parikh, A., Chang, M.-W., Das, D., and Cohen, W. W. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*, 2019.

Gardent, C., Shimorina, A., Narayan, S., and Perez-Beltrachini, L. Creating training corpora for nlg micro-planning. 2017a.

Gardent, C., Shimorina, A., Narayan, S., and Perez-Beltrachini, L. The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pp. 124–133, 2017b.

Goodrich, B., Rao, V., Liu, P. J., and Saleh, M. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 166–175. ACM, 2019.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Koncel-Kedziorski, R., Bekal, D., Luan, Y., Lapata, M., and Hajishirzi, H. Text generation from knowledge graphs with graph transformers. *arXiv preprint arXiv:1904.02342*, 2019.

Lebret, R., Grangier, D., and Auli, M. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*, 2016.

Luan, Y., He, L., Ostendorf, M., and Hajishirzi, H. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*, 2018.

Luong, M.-T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Marcheggiani, D. and Perez-Beltrachini, L. Deep graph convolutional encoders for structured data to text generation. *arXiv preprint arXiv:1810.09995*, 2018.

Marcheggiani, D. and Titov, I. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*, 2017.

Novikova, J., Dušek, O., and Rieser, V. The e2e dataset: New challenges for end-to-end generation. *arXiv preprint arXiv:1706.09254*, 2017.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Schmitt, M. and Schütze, H. Unsupervised text generation from structured data. *arXiv preprint arXiv:1904.09447*, 2019.

Sun, M., Li, X., Wang, X., Fan, M., Feng, Y., and Li, P. Logician: A unified end-to-end neural approach for open-domain information extraction. In *WSDM*, 2018.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Vougiouklis, P., Elsahar, H., Kaffee, L.-A., Gravier, C., Laforest, F., Hare, J., and Simperl, E. Neural wikipedian: Generating textual summaries from knowledge base triples. *Journal of Web Semantics*, 52:1–15, 2018.

Wiseman, S., Shieber, S. M., and Rush, A. M. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*, 2017.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *ArXiv*, abs/1906.08237, 2019.